Bohatá anotace ve frázovém strojovém překladu *

Aleš Tamchyna, Ondřej Bojar

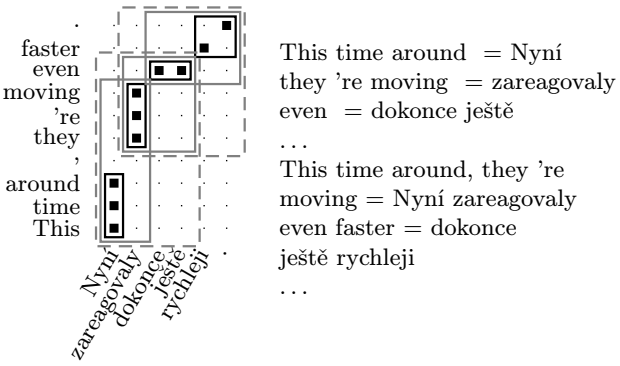
Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky
Malostranské nám. 25, Praha 1, CZ-118 00, Česká republika
A.Tamchyna@gmail.com, Ondrej.Bojar@mff.cuni.cz

Abstrakt *Mezi nejúspěšnější metody strojového překladu se v současné době řadí relativně velmi jednoduchý frázový statistický překlad, který se opírá v podstatě pouze o posloupnosti slov bez ohledu na lingvistické rozbory. Z více důvodů kvalita strojového překladu stále není uspokojivá a lze se domnívat, že část problémů by bylo možné odstranit explicitním zapojením lingvistické anotace do frázového překladu. Pro češtinu a angličtinu jsou navíc rozsáhlá bohatě anotovaná paralelní data k dispozici. Cílem této práce je proto připravit nástroj usnadňující experimenty s bohatou lingvistickou anotací v relativně jednoduchém prostředí frázových statistických překladů. Popisujeme formát dat i možnosti implementovaného nástroje a současně uvádíme výsledky prvních experimentů. Široký prostor možností, jak lingvistická data do modelu zapojit, je otevřen pro další výzkum.*

1 Úvod

Frázový statistický překlad (viz např. Koehn, 2009) je v současné době poměrně oblíbenou metodou strojového překladu. Jedním z důvodů je jednoduchost a velmi malá závislost metody na překládaných jazycích, dalším důvodem je bezesporu i existence volně šiřitelného překladače Moses založeného na tomto principu (Koehn et al., 2007).

Frázový překlad lze velmi stručně shrnout takto: texty dostupné současně ve zdrojovém a cílovém jazyce se tzv. zarovnají k sobě, tj. přiřadí se k sobě jednotlivé věty a v rámci dvojic vět i přibližně jednotlivá slova, viz obr. 1. Z takto zarovnaných vět extrahujeme všechny „fráze“, tj. posloupnosti slov, které jsou nějakým způsobem v souladu se slovním zarovnáním. Získaná **tabulka frází** (tj. slovník frází a jejich překladů) je pak použita i pro vstupní překládanou větu: vstupní věta se všemi možnými způsoby rozdělí na fráze a pro každou frázi se uváží všechny možné překlady. Výsledný strojový překlad se získá poskládáním přeložených frází za sebe tak, aby celá vstupní věta byla přesně pokryta a výsledná sekvence navíc působila „hladce“, tj. získala co největší skóre v tzv. jazykovém modelu. Kandidátů na výsledné sekvence je přirozeně obrovské množství, a v praxi je proto tento prostor prohledáván jen částečně metodou beam search.



Obrázek 1. Ukázkové zarovnání po slovech a příklady „frází“ konzistentních s daným zarovnáním. (V obrázku nejsou vyznačeny všechny možné fráze.)

Je zřejmé, že tento jednoduchý postup v řadě případů navrhne výstup s mnoha chybami. Kromě chybné volby překladového ekvivalentu je velmi často výstup negramatický, a to jak lokálně (např. shoda mezi přídavným a podstatným jménem), tak na úrovni celé věty (ve větě např. zcela chybí sloveso).

Lokální problémy je možné často řešit v rámci frázového modelu rozšířením reprezentace jednotlivých frází. Zavedeným způsobem, který je rovněž implementován v překladači Moses, jsou tzv. frázové modely o více faktorech (Koehn – Hoang, 2007). Data se v tomto případě získávají z anotovaného korpusu – ke slovům ve větách se doplní tzv. faktory, tedy značky, které určují např. slovní druh nebo funkci slova ve větě. Tato dodatečná data se zahrnou i do frázové tabulky a dekodér získá možnost vytvořit přesnější překlad, viz obr. 2.

V tomto příspěvku popisujeme nástroj, který z bohatě anotovaného korpusu zejména připraví podle konfigurace výběr některých rysů slov ve větě do formátu vhodného pro překladový systém Moses. Důraz je kladen právě na stručnou a příp. i strojově generovatelnou konfiguraci, aby bylo možné v krátkém čase provést celou řadu experimentů. Bez experimentálního ově-

* Tato práce je podporována granty EuroMatrixPlus (FP7-ICT-2007-3-231720 EU a 7E09003 České republiky) a MSM 0021620838.

Základní překlad

[**Peter**] [slept] →[**Petr**] [spal]
[I saw] [**Peter**] →[Viděl jsem] [**Petra**]

Vícefaktorový překlad

$$\begin{bmatrix}\textbf{Peter}\\ \text{subj}\end{bmatrix}\begin{bmatrix}\text{slept}\\ \text{verb}\end{bmatrix}\rightarrow\begin{bmatrix}\textbf{Petr}\\ \text{nom.}\end{bmatrix}\begin{bmatrix}\text{spal}\\ \text{slov.}\end{bmatrix}$$

$$\begin{bmatrix}\text{I saw}\\ \text{vp}\end{bmatrix}\begin{bmatrix}\textbf{Peter}\\ \text{obj}\end{bmatrix}\rightarrow\begin{bmatrix}\text{Viděl jsem}\\ \text{slov. fr.}\end{bmatrix}\begin{bmatrix}\textbf{Petra}\\ \text{akuz.}\end{bmatrix}$$

Obrázek 2. Bohatší reprezentace ve frázovém přístupu umožní překládat anglické slovo „Peter" do vhodného českého pádu podle pozice, kterou v původní větě zaujímalo, protože vícefaktorová frázová tabulka nově tyto dvě role slova Peter pokládá za zcela odlišná vstupní slova.

ření nelze totiž říci, které z dostupných lingvistických údajů samotné úloze překladu pomohou.

1.1 Předešlé práce

V (Avramidis – Koehn, 2008) autoři uvádějí dosažení lepší kvality překladu díky použití faktorů popisujících syntax věty. Tato metoda se nejvíce osvědčila při překladu mezi morfologicky bohatými a chudými jazyky, kde faktory pomohly rozlišit, který tvar slova se má použít. Birch et al. (2007) využívají tzv. CCG supertags (značky kategoriální gramatiky) ke zlepšení překladu v situacích, kdy je nutné výrazně měnit pořádek slov ve větě.

Dalším možným přístupem k překladu s využitím bohaté anotace jsou různé druhy tvorby frází a jejich filtrování. Máme-li k dispozici např. syntaktický závislostní strom věty, můžeme fráze vytvářet tak, aby byly v souladu se skutečnou syntaxí věty. Případně je možné filtrovat takové fráze, které jsou se závislostním stromem nekonzistentní, jak navrhuje Bojar (2009).

U všech uvedených metod je klíčovým pojmem (bohatě) anotovaný korpus. Ten obsahuje věty v požadovaných jazycích obohacené o mnoho dalších informací. Pro další popis se přidržíme formalismu jazykového popisu na více rovinách, který byl pro češtinu rozpracován v lingvistické teorii Funkčního generativního popisu (FGP, Sgall et al., 1986) a následně použit při anotaci Pražského závislostního korpusu (Prague Dependency Treebank, PDT, Hajič, 2004). V současné době jsou stejné principy uplatňovány i pro angličtinu v paralelním korpusu Prague Czech-English Dependency Treebank (verze 1.0 viz Čmejrek et al. (2004), verze 2.0 je v přípravě) a odrážejí se i v nástrojích pro automatickou lingvistickou analýzu pro češtinu, angličtinu a pracovně i pro další jazyky, viz TectoMT (Žabokrtský – Bojar, 2008).

Anotace v trénovacím paralelním korpusu tedy může být dostupná na několika rovinách. Na morfologické

- **anglická a-rovina:**
  `Mae|Mae|NNP|1|3|Sb`
  `just|just|RB|2|3|Adv left|leave|VBD|3|0|Pred`
  `.|.|.|4|0|AuxK`
- **anglická t-rovina:**
  `Mae|ACT|1|3|complex|n:subj|n.denot`
  `just|RHEM|2|3|atom|x|-`
  `leave|PRED|3|0|complex|v:fin|v`
- **mapování anglická a-rovina → anglická t-rovina:** `0-0 1-1 2-2`
- **česká a-rovina:**
  `Mae|Mae|X@---|1|3|Adv`
  `právě|právě|Db---|2|3|Adv`
  `odešla|odejít|VpQW-|3|0|Pred`
  `.|.|Z:---|4|0|AuxK`
- **česká t-rovina:**
  `mae|TWHEN|1|4|complex|n:???|n.denot`
  `právě|TWHEN|2|4|complex|adv:|adv.denot.ngrad.nneg`
  `#PersPron|ACT|3|4|complex|n:1|n.pron.def.pers`
  `odejít|PRED|4|0|complex|v:fin|v`
- **mapování česká a-rovina → česká t-rovina:** `0-0 1-1 2-3`
- **mapování mezi anglickou a českou t-rovinou:** `0-0 1-1 2-3`

Obrázek 3. Exportní formát CzEngu. Reprezentace věty *Mae just left.* (*Mae právě odešla.*). Některé prvky byly pro lepší čitelnost zjednodušeny nebo vypuštěny.

rovině jsou slova zařazena do tvaroslovných kategorií, analytická rovina rozebírá tzv. povrchovou syntax věty (vztahy typu podmět – přísudek apod.). Nejkomplexnější anotace se (v teorii FGP a pražských treebancích) odehrává na tektogramatické rovině, kde se sledují hlubší vztahy mezi slovy, hloubková syntax a další. Je zřejmé, že bohatá anotace má mnoho možností využití a najít vhodnou kombinaci faktorů pro zlepšení překladu je složité.

Využívání bohaté anotace při frázovém strojovém překladu se v současné době prosazuje a díky zmíněným výsledkům se dá očekávat, že experimentů s tímto přístupem bude přibývat.

1.2 Dostupná data

CzEng (Bojar – Žabokrtský, 2009) je bohatě anotovaný česko-anglický paralelní korpus. V současné verzi 0.9 obsahuje přibližně 8 milionů vět. Data byla získána z mnoha zdrojů, např. filmové titulky, legislativa Evropské unie, beletrie, technické texty a další. Paralelní data byla automaticky anotována na několika rovinách – morfologické, analytické a tektogramatické.

Exportní formát CzEngu popisuje bohatou anotaci úsporněji než běžně používané XML. Jedná se o textový formát, každá dvojice vět je zapsána na jednom řádku do sloupců oddělených znakem tabulátoru. Strukturu řádky popíšeme na příkladu v obrázku 3.

Vzhledem k tomu, že uzly na morfologické a analytické rovině si navzájem odpovídají, jsou v exportním formátu v zájmu úspory tyto roviny sloučeny do jedné, kterou označujeme jako analytickou. Tektogramatická rovina obsahuje pouze plnovýznamová slova (až na výjimky), oproti analytické rovině v ní mohou naopak přebývat slova, která jsou ve větě nevyjádřená. Uzly této roviny jsou tedy popsány zvlášť.

Ze všech údajů dostupných v korpusu CzEng zde popišme jen několik nejvýznamnějších:

form obsahuje povrchový tvar slova, tzv. formu.

lemma a tlemma obsahuje základní tvar slova po odstranění morfologické informace.

tag nese morfologickou značku, která (s ohledem na konkrétní jazyk) obsahuje kompaktní přehled o hodnotách lingvistických kategorií pro dané slovo. Např. české „NNMS4“ signalizuje podstatné jméno (NN) mužského rodu (M) jednotného čísla (S) ve čtvrtém pádě.

functor na tektogramatické rovině vyjadřuje syntaktickou roli daného uzlu ve větě (např. PRED pro predikát, RSTR pro přívlastek).

formeme na tektogramatické rovině vyjadřuje povrchovou formu realizace daného uzlu ve větě (např. „v:fin“ pro sloveso v určitém tvaru nebo „n:na+6“ pro podstatné jméno v šestém pádě s předložkou „na“).

sempos vyjadřuje sémantický slovní druh tektogramatického uzlu. Např. přídavná jména mohou stát v roli podstatných jmen, pokud žádné podstatné jméno nerozvíjejí.

nonterm v anglické povrchové syntaxi vyjadřuje syntaktickou roli skupiny slov ve větě (např. NP značí jmennou skupinu).

2 Funkce nástroje Richextr

Hlavním účelem Richextru je široce konfigurovatelné, automatické předzpracování vstupních dat z bohatě anotovaného korpusu. Jeho další funkcí je extrakce frázových tabulek s možností označování frází podle konzistence s kontrolními zarovnáními slov. Richextr také umožňuje provádět na frázových tabulkách základní množinové operace.

2.1 Zpracování bohatě anotovaných dat

Richextr umožňuje uživateli velmi volně konfigurovat formát vstupních dat (tj. anotovaného korpusu), není závislý na konkrétním počtu rovin anotace, mapování mezi nimi, počtu ani druhu faktorů. Ačkoliv jsme tedy Richextr využívali především pro práci s daty v exportním formátu CzEngu, změnou v konfiguračním souboru lze Richextr použít na korpusy s odlišným formátem.

Nejpodstatnější vlastnost při zpracování bohatě anotovaného korpusu je konfigurovatelnost požadovaných výstupních dat. Richextr interně uchovává vstupní data ve formě závislostních stromů na jednotlivých rovinách. Tento přístup umožňuje programu tyto stromy procházet a získávat faktory z jiných uzlů stromu, případně (je-li k dispozici mapování) z uzlů na jiných rovinách anotace.

Uživatel proto musí určit cestu, kterou má Richextr projít, a požadované faktory. Jednotlivé položky jsou v konfiguraci oddělovány znakem lomítka. Richextr rozpoznává klíčová slova *parent*, *son*, *son_with*, *link* a *factor*. První určuje přechod na otcovský uzel, druhé na synovský uzel (se zadaným indexem), třetí umožňuje specifikovat synovský uzel názvem a hodnotou faktoru. Klíčové slovo *link* způsobí přechod na zadanou rovinu anotace (mezi aktuální a cílovou rovinou musí existovat mapování). Slovem *factor* musí končit každý popis cesty – určuje název faktoru, jehož hodnota se má přidat do výstupu. Např. popis

```
parent/link(tLayerCS)/factor(tlemma)
```

definuje následující cestu po stromech věty: z uzlu na aktuální rovině přejdi na otce, z něj přejdi na rovinu *tLayerCS* a odtud přečti faktor *tlemma*.

2.2 Extrakce frázových tabulek

Extrakce frázových tabulek přirozeně navazuje na předzpracování korpusu. V situaci, kdy Richextr získal požadovaná data, je možné (je-li k dispozici zarovnání slov) vytvářet frázovou tabulku. Díky informacím o mapování mezi jednotlivými rovinami anotace lze navíc kontrolovat konzistenci frází např. s tektogramatickým zarovnáním už v průběhu jejich vytváření.

Richextr zároveň umožňuje kontrolovat konzistenci fráze se zarovnáními uloženými v souborech.

2.3 Množinové operace

Další funkcí programu Richextr jsou operace sjednocení a průniku na setříděných frázových tabulkách. Richextr pracuje se standardním formátem frázové tabulky používaným v dekodéru Moses. Množinové operace dovoluje provádět na libovolném počtu souborů s frázovými tabulkami, podporuje přepočítání hodnot pravděpodobností a lexikálních vah podle zadaných koeficientů pro jednotlivé tabulky.

3 První experimenty

V této sekci podrobněji popíšeme postup a výsledky provedených experimentů a pokusíme se naznačit mož-

né cesty, jak pomocí bohaté anotace zlepšit kvalitu strojového překladu.

3.1 Použité nástroje

K provedení experimentů s bohatou anotací jsme kromě programu Richextr využili řadu dalších lingvistických nástrojů. Všechny použité programy mají otevřený zdrojový kód a jsou volně dostupné.

Pro tvorbu jazykového modelu jsme využili nástroj SRILM[1], k nalezení slovních zarovnání program Giza++[2]. K samotnému překladu jsme využili zmiňovaný nástroj Moses, který obsahuje také řadu pomocných skriptů a menších programů pro přípravu dat, práci s frázovými tabulkami, hodnocení kvality překladu apod., které jsme v práci využívali. Jeho součástí je i nástroj MERT, s jehož pomocí jsme prováděli ladění vah.

Při hodnocení kvality překladu jsme se spoléhali na běžně používanou automatickou metriku BLEU (Papineni et al., 2002).

3.2 Vstupní data

Vytvořit bohatě anotovaný paralelní korpus je velmi náročné, vhodná vstupní data jsou často obtížně dostupná. V experimentech jsme se proto omezili na rozsáhlý korpus CzEng a tím na překlad mezi dvěma jazyky – angličtinou a češtinou. Jako formát vstupních dat jsme použili výše popsaný exportní formát tohoto korpusu.

Aby bylo možné otestovat co nejvíce konfigurací a zároveň tak ověřit funkčnost programu Richextr, použili jsme na experimenty jen malý vzorek dat – paralelní korpus o velikosti 30 tisíc vět, dalších 300 vět k ladění a 300 k ověření kvality překladu. Výsledky experimentů jsou proto pouze orientační, pro jejich ověření by bylo nutné provést je na několikanásobně větších datech.

3.3 Scénáře experimentů

Kombinace faktorů. V těchto experimentech jsme se zaměřili na přímočaré využití nástroje Richextr – extrakci různých faktorů ze závislostních stromů a tvorbu nového korpusu. Zahrnutí dalších faktorů by mohlo zlepšit kvalitu překladu.

Ve všech experimentech byla použita stejná vstupní data, bylo tedy možné připravit slovní zarovnání předem. Pomocí programu Richextr jsme provedli lemmatizaci dat (extrakcí lemmat z korpusu), nástroj Giza++ jsme spustili oběma směry a provedli symetrizaci zarovnání algoritmem *grow-diag-final*.

[1] http://www-speech.sri.com/projects/srilm/

[2] http://fjoch.com/GIZA++.html

Vstupní data (exportní formát CzEngu) jsme spolu s konfigurací pro daný experiment předložili na vstup programu Richextr, který vytvořil nový paralelní korpus s požadovanými faktory. Ten jsme převedli na malá písmena a vytvořili jazykový model cílového jazyka řádu 3 s parametry *-interpolate -kndiscount*.

Jazykový model, slovní zarovnání a vytvořený korpus jsme použili k trénování překladového modelu pomocí skriptu `train-model.perl`. Tento skript poskytuje možnost automatizovat velkou část experimentu, my jsme využili ty části, které jsme neřešili pomocí vlastních nástrojů. Po dokončení skriptu je již vytvořena frázová tabulka a další pomocné modely. Následně jsme provedli ladění překladového modelu (tato část experimentu je časově nejnáročnější, i na velmi malých datech může trvat několik hodin).

Po dokončení ladění jsme přeložili dosud neviděná data a pomocí skriptu `multi-bleu.pl` porovnali výsledek s referenčním překladem, abychom zjistili skóre BLEU. Převod *tokenů* zpět na věty jsme vynechali, neboť i referenční překlad byl tokenizován.

Vážené sjednocení. Richextr poskytuje při operacích sjednocení a průniku frázových tabulek možnost přepočítat skóre frází podle zadaných vah. Tuto vlastnost jsme se pokusili využít pro zvýhodnění těch domén textu, které lze považovat za kvalitně přeložené, tj. sekce *news* (zpravodajství) a *fiction* (beletrie).

Experimenty jsme provedli pouze s operací sjednocení, neboť po vyloučení ostatních frází, ke kterému by došlo při průniku, by ve frázové tabulce zůstaly pouze věty ze zvýhodněných domén.

Při experimentech jsme využívali stejné nástroje jako v předchozím scénáři, postup ovšem bylo nutné pozměnit. Kromě extrakce frázové tabulky pro překlad jsme provedli také extrakci frází pouze z domén *news* a *fiction*. Obě frázové tabulky jsme předali programu Richextr, který vytvořil jejich vážené sjednocení. Tím jsme nahradili původní frázovou tabulku a nechali proces trénování překladového modelu pokračovat.

Poté jsme provedli ladění modelu a testování kvality překladu stejným způsobem jako v předchozím scénáři.

Konzistence frází. Abychom ověřili funkčnost extrakce frází a kontroly jejich konzistence, vytvořili jsme scénář experimentu, ve kterém Richextr tyto úlohy provede. Do frázových tabulek je přidána informace o konzistenci každé fráze se zvolenými dodatečným mapováním v podobě čísla 1 nebo 0. Překladač by měl následně upřednostňovat fráze, které jsou označeny jako konzistentní s dodatečným zarovnáním, protože ty budou mít obvykle vyšší skóre.

Integrace vlastní neohodnocené frázové tabulky a zahrnutí sloupce popisujícího konzistenci frází do trénování překladového modelu vyžadovalo několik úprav dosavadního postupu.

Skript, který trénování modelu zajišťuje, jsme spouštěli po malých krocích a postupně jsme vkládali vlastní data. Krok extrakce frází jsme nahradili spuštěním programu Richextr a po ohodnocení frází jsme k výsledné frázové tabulce připojili sloupec s informací o konzistenci fráze.

Při ladění tohoto překladového modelu pak nástroj MERT bral v úvahu i námi přidaný příznak 0/1.

Výsledek učení jsme opět ověřili testovacím překladem.

3.4 Překlad do češtiny

Překlad z angličtiny do češtiny je poměrně obtížný. Jedním z hlavních důvodů je velmi bohaté tvarosloví češtiny. Stejná slova a fráze v angličtině se mohou do češtiny překládat různě v závislosti na kontextu (např. tvar slovesa ve větě závisí na podmětu). Pro dekodér je proto v podstatě nemožné bez jiných informací mimo základní frázové tabulky větu správně přeložit.

Automatické metriky, které ověřují kvalitu překladu a při velkých sadách experimentů je nutné se na ně alespoň částečně spoléhat, navíc správné slovo v chybném tvaru považují za chybu (nijak nerozlišují morfologii), čímž se snižuje skóre takového překladu a tím i možnosti rozlišení srozumitelných nesprávných překladů od překladů zcela chybných.

S využitím bohaté anotace jsme se proto pokusili najít vhodné kombinace faktorů, které by pomohly překonat nedostatek informací při překladu. Abychom dokázali rozlišit vhodné faktory, zvolili jsme co nejjednodušší scénář překladu – kombinujeme vždy pouze dva faktory ve zdrojovém jazyce, v cílovém jazyce překládáme pouze do formy, zahrnujeme vždy také možnost překladu pouze pomocí prvních faktorů v každém jazyce pro překonání nevyhnutelné řídkosti dat.

Následující tabulka shrnuje dosažené výsledky:

Faktory	BLEU
form	14,19
form, link(ent)/tlemma	14,21
form, link(ent)/functor	14,15
form, tag	13,64
form, parent/link(ent)/functor	13,69
form, nonterm	13,96
form, link(ent)/formeme	**15,14**
form, link(ent)/sempos	14,14
form, link(ent)/parent/son(0)/tlemma	14,33

Všechny výsledky se pohybovaly kolem hranice 14 BLEU, některé kombinace byly dokonce horší než základní překlad forma→forma. Vzhledem k malým rozdílům lze tyto hodnoty pravděpodobně přičíst náhodnému postupu při ladění modelu metodou MERT.

Většina testovaných kombinací faktorů se tedy jeví jako nepříliš užitečná. Tento výsledek je nejspíše způsoben faktem, že žádný z dodatečných faktorů nepřináší dostatečně podrobné informace pro rozlišení mezi možnými překlady do češtiny, např. přidání morfologické značky anglického slova jen málo vypovídá o tom, jaký tvar by mělo mít české slovo v překladu. Poněkud překvapivé je však zjištění, že tyto dodatečné informace nepomáhají kvalitě překladu vůbec.

Jediným výrazným výsledkem je překlad obohacený o faktor *formeme* z tektogramatické roviny. Tento faktor reprezentuje kombinaci morfologického a syntaktického popisu uzlu (Žabokrtský et al., 2008). Skóre téměř o 1 BLEU vyšší než u obyčejného překladu naznačuje, že by jeho využití mohlo skutečně přinést zlepšení kvality překladu.

Pro experimenty s váženým sjednocením jsme zvolili pouze základní překlad mezi formami, jehož výsledky lze snadno interpretovat. Váhy vět z kvalitnějších sekcí jsme volili ručně, váha ostatních sekcí byla vždy 1. V tabulce jsou shrnuta dosažená skóre:

Váha	BLEU
2	**14,34**
4	14,18

Nejvyšší dosažené zlepšení překladu je velmi malé (0,15 BLEU). Přehnané zvýhodnění kvalitnějších textů má na kvalitu překladu negativní vliv. Nelze s určitostí rozhodnout, zda je popsané mírné navýšení kvality náhodným výsledkem. Bylo by nutné provést experimenty s rozsáhlejšími trénovacími daty a větším počtem kombinací vah.

Scénář s kontrolou konzistence frází jsme využili pouze pro jeden experiment. Do překladového modelu jsme zahrnuli informaci o konzistenci s tektogramatickou rovinou, kterou jsme získali nástrojem Richextr. Dosáhli jsme výsledku pouhých 9,60 BLEU, tedy o několik bodů horšího než u předchozích experimentů.

Příčina tohoto skóre není jasná, nejspíše se zde negativně projevil způsob, jakým Richextr extrahuje fráze (ačkoliv se jeho výstup z velké části shoduje se standardním nástrojem pro extrakci). Určitý vliv mělo pravděpodobně i ladění překladového modelu, při kterém jsme dovolili přidělení velké váhy informaci o konzistenci. To mohlo způsobit upřednostňování frází, které byly sice konzistentní, zato však málo pravděpodobné.

3.5 Překlad do angličtiny

Tento směr překladu obvykle dává mnohem lepší výsledky (Koehn et al., 2006). Jedním z důvodů tohoto

rozdílu je fakt, že angličtina má velmi jednoduchou morfologii, a tak se překlad vyhýbá chybám ve tvarech slov.

Nastává zde ovšem problém řídkosti dat, protože ačkoliv se různé tvary slova v češtině většinou mají přeložit na stejný tvar anglického slova, dekodér musí daný zdrojový slovní tvar najít ve frázové tabulce. Pokud se tento tvar v trénovacích datech nevyskytl, překladač vydá nepřeložené slovo. Experimenty jsme proto kromě testování slibných kombinací faktorů zaměřili také na překonání této překážky. V cílovém jazyce jsme opět použili pouze faktor *form*. Dosažené výsledky popisuje následující tabulka:

Faktory	BLEU
form	18,65
form, link(cst)/formeme	18,27
form, tag	18,91
lemma, tag	**20,04**
lemma	19,08

V tomto směru byla skóre překladů podle očekávání vyšší, pohybovala se okolo hodnoty 18,5 BLEU. Stejně tak se potvrdil předpoklad, že pokusíme-li se vyhnout řídkosti dat, dosáhneme vyššího skóre – kombinace lemmatu a morfologické značky přinesla znatelně lepší výsledek než ostatní experimenty. Slov v základním tvaru je zřejmě mnohem méně a informace o tvaru slova dává dekodéru možnost přesnějšího rozhodování než jednoduchý překlad lemma→forma.

Skóre varianty s faktorem *lemma* bylo poměrně překvapivé. Vzhledem k tomu, že data o češtině byla velmi zjednodušena, dal se očekávat spíše propad způsobený neschopností překladače správně rozlišit mezi variantami překladu stejné kombinace lemmat. Tento problém je podobný při překladu v opačném směru – nedostatek informací ve zdrojovém jazyce komplikuje výběr správného překladu. Fakt, že kvalita překladu naopak vzrostla, lze přičíst malé velikosti trénovacích dat, kvůli které se výrazně projevila řídkost dat.

Stejně jako v opačném směru překladu jsme i zde provedli dva experimenty s využitím váženého sjednocení. Výsledky popisuje tabulka:

Váha	BLEU
2	**18,94**
4	18,20

Přestože jsou data velmi malá, takže nelze přímo usuzovat na praktický přínos tohoto postupu, projevuje se určité navýšení kvality překladu (při váze 2 o 0,29 BLEU), navíc podobné výsledkům dosaženým v opačném směru. Dá se proto předpokládat, že zvýhodnění textů, u kterých je větší jistota správnosti, mírně zlepšuje kvalitu překladu.

Provedli jsme jeden experiment s využitím kontroly konzistence frází. Jeho výsledek 15,05 BLEU potvrzuje, že tento scénář je náročnější na správné provedení. Nízké skóre lze, podobně jako při překladu do češtiny, patrně přičíst špatné konfiguraci ladicího nástroje a odlišnostem ve výstupu extrakce frází programu Richextr a standardního nástroje.

4 Závěr a budoucí pokusy

Představili jsme nástroj Richextr pro výběr lingvistické anotace z bohatě anotovaného paralelního korpusu. Nástroj je široce konfigurovatelný a nabízí tak celou řadu možností, která konkrétní data v překladu zkusit použít.

První experimenty na velmi malé sadě trénovacích vět naznačují, že např. tzv. formém z hloubkové anotace angličtiny pomáhá správně generovat český tvar při překladu do češtiny. Pro opačný směr se naopak jako vhodné jeví opřít se o základní tvar českého slova a tím omezit nedostatečnou velikost trénovacích dat, kde mnohé tvary českých slov nebyly nikdy spatřeny.

V budoucí práci se zaměříme na automatizaci celého postupu: budeme strojově generovat nové konfigurace rozvíjením těch dosud nejnadějnějších a ověřovat je v praxi. Věříme, že se tak podaří využít potenciál dostupného bohatě anotovaného korpusu a zlepšit kvalitu překladu zejména ve směru z angličtiny do češtiny. Implementovaný Richextr je však samozřejmě naprosto jazykově nezávislý.

Literatura

Avramidis, E., Koehn, P. *Enriching Morphologically Poor Languages for Statistical Machine Translation.* In Proceedings of ACL-08: HLT, s. 763–770, Columbus, Ohio, 2008. Association for Computational Linguistics. Dostupné z: http://www.aclweb.org/anthology/P/P08/P08-1087.

Birch, A., Osborne, M., Koehn, P. *CCG Supertags in Factored Statistical Machine Translation.* In Proceedings of the Second Workshop on Statistical Machine Translation, s. 9–16, Praha, 2007. Association for Computational Linguistics. Dostupné z: http://www.aclweb.org/anthology/W/W07/W07-0202.

Bojar, O. *Bad News, NLP Hacking and Feature Fishing.* MT Marathon 2009, Praha, 2009.

Bojar, O., Žabokrtský, Z. *CzEng0.9: Large Parallel Treebank with Rich Annotation.* Praha. 2009, 92. ISSN 0032-6585.

Čmejrek, M. a kol. *Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation.* In Proceedings of LREC 2004, Lisbon, May 26–28 2004.

Hajič, J. *Complex Corpus Annotation: The Prague Dependency Treebank.* In Insight into Slovak and Czech Corpus Linguistics, Bratislava, Slovakia, 2004. Jazykovedný ústav Ľ. Štúra, SAV. ISBN 80-224-0880-8.

Koehn, P. Statistical Machine Translation. : Cambridge University Press, 2009.

Koehn, P., Hoang, H. *Factored Translation Models.* In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), s. 868–876, Praha, 2007. Association for Computational Linguistics. Dostupné z: http://www.aclweb.org/anthology/D/D07/D07-1091.

Koehn, P. a kol. *Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding.* Technical report, Johns Hopkins University, Center for Speech and Language Processing, 2006.

Koehn, P. a kol. *Moses: Open Source Toolkit for Statistical Machine Translation.* In ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, s. 177–180, Praha, 2007. Association for Computational Linguistics. Dostupné z: http://www.aclweb.org/anthology/P/P07/P07-2045.

Papineni, K. a kol. *Bleu: a Method for Automatic Evaluation of Machine Translation.* In ACL, s. 311–318, 2002. Dostupné z: http://www.aclweb.org/anthology/P02-1040.pdf.

Sgall, P., Hajičová, E., Panevová, J. The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Prague, Czech Republic/Dordrecht, Netherlands : Academia/Reidel Publishing Company, 1986.

Žabokrtský, Z., Bojar, O. *TectoMT, Developer's Guide.* Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, December 2008.

Žabokrtský, Z., Ptáček, J., Pajas, P. *TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer.* In ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation, s. 167–170, 2008.